Comparing Automated Factual Claim Detection Against Judgments of Journalism Organizations

Naeemul Hassan[§], Mark Tremayne[¶], Fatma Arslan[§], Chengkai Li[§]

[§]Department of Computer Science and Engineering
[¶]Department of Communication
The University of Texas at Arlington

ABSTRACT

In this research, we deployed an automated fact-checking system (ClaimBuster) on the 2016 presidential primary debates and assessed its performance compared to professional news organizations. In real-time, ClaimBuster scored the statements made by candidates on check-worthiness, the likelihood that the statement contained an important factual claim. Its discrimination compared to professional journalists was high. Statements chosen for fact checking by CNN and PolitiFact had been scored much higher by ClaimBuster than those not selected. The topics of statements chosen for checking also mirrored the topics of those scored highly by ClaimBuster. Differences between the political parties and individual candidates on the use of factual claims are also presented.

1. INTRODUCTION

Real-time fact checking of important events like political debates remains a challenge for those pursuing it. The first step is to identify the important claims - those that are check-worthy. In this paper, we use all the 21 primary debates for the 2016 U.S. presidential election to compare claims picked by ClaimBuster - our machine-learning based algorithm and system - against the judgments of professional fact checkers at CNN and PolitiFact.

While fact checking has long been a staple of journalism, a competing norm, objectivity, led some news organizations to shy away from contradicting claims made by partisans [9]. The last decade has seen renewed interest in fact checking [11]. Recent research suggests this may have less to do with audience demand than it does with status achievement within the field of journalism [4]. Whatever the cause, research has demonstrated the benefits for news consumers. A study on the 2012 presidential race found that those who visited a fact checking website like PolitiFact were better informed about the race than those who did not [3]. Another study found that consumption of fact checking articles influenced how citizens evaluated negative political ads [2]. A study also found that politicians informed of the electoral and reputational consequences of being fact checked were significantly less likely to receive a negative fact-check rating or have their accuracy publicly challenged [8].

The benefits of fact checking are clear but the number of news organizations with the resources to do it are few. A recent study deemed the practice "quite rare" and found only 3 percent of reporters surveyed using it in their stories [4]. Automated fact-checking tools can be part of the solution [1, 12, 6]. Separating factual claims from opinions and further discriminating between important and unimportant factual claims is our first step in building ClaimBuster. We used it to find important factual claims in the aforementioned 21 primary debates. We also performed topic-detection on the debate transcripts and studied the use of factual claims on the different topics addressed by the Republican and Democratic presidential candidates. We coupled this data with fact checks performed by CNN[1] and PolitiFact[2] and then compared the performance of ClaimBuster against these professional organizations.

2. OVERVIEW OF CLAIMBUSTER

ClaimBuster [3] [6, 5] is a tool that helps journalists find claims to fact-check. While its details can be found in [6], we provide a brief overview of the system in this section.

Given a sentence, ClaimBuster uses a classification and ranking model to determine how check-worthy the sentence is. The model was trained over human-annotated 1960-2012 general election debate transcripts. It uses the tokens in sentences and the tokens' part-of-speech (POS) tags as the features in the model. ClaimBuster assigns a score between 0.0 and 1.0 to each sentence. The higher the score, the more likely the sentence contains check-worthy factual claims. The lower the score, the more non-factual, subjective and opinionated the sentence is.

Figure 1 is a screenshot of ClaimBuster when it is applied on a debate. The background colors of the sentences indicate how check-worthy they are. Darker colors correspond to higher check-worthiness scores. By default, all sentences having scores higher than or equal to 0.5 are highlighted. A slider allows the user to modify this threshold. An *Order by Score* button allows the user to order all the sentences by their check-worthiness scores. This way of ranking helps fact-checkers prioritize their efforts in assessing the veracity

[1]www.cnn.com
[2]www.politifact.com
[3]idir.uta.edu/claimbuster

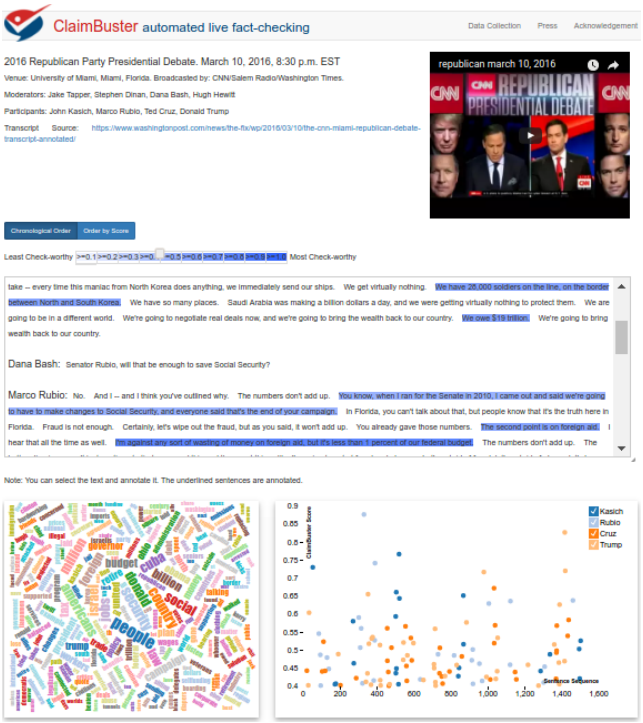

Figure 1: ClaimBuster platform

of claims. Thus, ClaimBuster will free journalists from the time-consuming task of finding check-worthy claims, leaving them with more time for reporting and writing.

3. PREPARATION FOR THE STUDY

Using the 21 primary debates, we compared ClaimBuster against the human fact-checkers at several popular fact-checking organizations. We are interested in testing the hypothesis that the claims picked by ClaimBuster are also more likely to be fact-checked by professionals. If the hypothesis is true, we can expect ClaimBuster to be effective in assisting professionals choose what to fact-check and thus helping improve their work efficiency.

3.1 Data Collection

There have been 12 Republican[4] and 9 Democratic primary debates in the 2016 U.S. presidential election. The debates featured as many as 11 Republican Party candidates and 5 Democratic Party candidates at the beginning, respectively. These debates took place between August, 2015 and April, 2016. We collected the transcripts of all these debates from several news media websites, including Washington Post, CNN, Times, and so on. There are a total of 30737 sentences in the 21 transcripts. We pre-processed these transcripts and identified the speaker of each sentence. Furthermore, we identified the role of the speaker. Sentences spoken by debate moderators were excluded from the study presented in this paper.

3.2 Finding Check-worthy Factual Claims

[4]We only considered the "prime time" debates which included the more popular candidates.

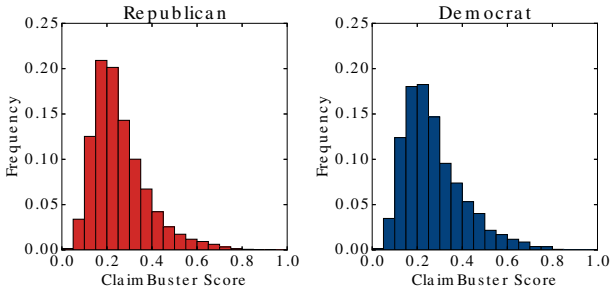

Figure 2: Distributions of ClaimBuster scores over all the sentences for both parties

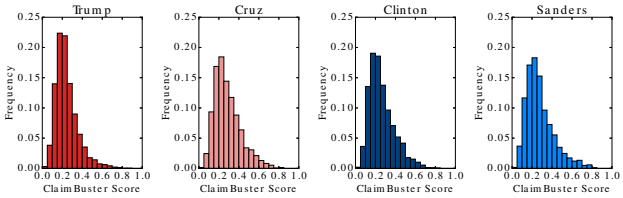

Figure 3: Distributions of ClaimBuster scores over all the sentences for the major candidates

We use ClaimBuster to calculate the check-worthiness scores of the sentences and thereby identify highly check-worthy factual claims. Figure 2 shows the distributions of ClaimBuster scores on all the sentences for both political parties. The distributions for the two parties are similar. One distinction is that the distribution for the Republican Party has a higher peak and a slightly thinner right tail than the distribution for the Democratic party. There are 776 check-worthy factual claims spoken by the Republicans with ClaimBuster scores over 0.5. This is 5.06% of all the sentences spoken by the Republican candidates. From Democrat candidates, there are 484 (6.73%) sentences with ClaimBuster score higher than 0.5.

Figure 3 shows the check-worthiness score distributions for the major candidates (nomination winners and runner-ups) from both parties. Among these four candidates, *Donald Trump* appears to have presented less number of highly check-worthy factual claims (ClaimBuster score ≥ 0.5) than the other three candidates. He has used more non-factual sentences (ClaimBuster score ≤ 0.3) compared to the other candidates.

3.3 Topic Detection

From each of the 21 debates, the 20 highest-scoring sentences were selected and manually placed in topic categories, a modified version of the most important problems (MIP) used by Gallup and other researchers for decades [7, 10, 13]. The major topics in the primary debates were: economy, crime, international affairs, immigration, health care, social issues, education, campaign finance, environment, Supreme Court, privacy and energy. Some of these topics were further broken down into subtopics. The 420 sample sentences were used to cultivate a list of keywords most often found for each of these topics. For example, the keywords for subtopic "abortion" were abortion, pregnancy and planned parent-

Platforms	avg(YES)	avg(NO)	t-value	p-value
CNN	0.433	0.258	21.137	1.815E-098
PolitiFact	0.438	0.258	16.362	6.303E-060

Table 1: Score differences between sentences fact-checked and those not chosen for checking

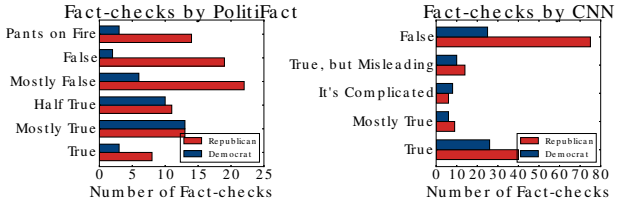

Figure 4: Distribution of verdicts for each party

hood. Some topics had a small number of keywords, others had more than 20.

A topic-detection program is created to detect each debate sentence's topic. Provided a sentence, the program computes a score for each topic in our list based on presence of each topic's keywords in the sentence. The score is the total number of occurrences of such keywords. The sentence is assigned to the topic attaining the highest score among all the topics. However, if the highest score is lower than a threshold (two occurrences of topic keywords), the program does not assign any of the topics to the sentence. If there is a tie between two or more topics, the program uses the topic of the preceding sentence if it matches one of the tied topics. Otherwise, it randomly picks one of the tied topics.

In order to evaluate the above approach to detect topics, we created ground-truth data for one Republican debate and one Democratic debate. We only used sentences with at least 0.5 ClaimBuster score. In our ground-truth data for the Democratic debate, there are 52 sentences and 39 of them are labeled with a topic. The program detected topics for 27 of the 39 sentences and only one sentence was assigned with a incorrect topic. For the Republican debate ground-truth data, there are 62 sentences and 44 sentences are labeled with a topic. The program found topics for 30 out of the 44 sentences and 5 of these sentences were mis-classified.

We applied the topic detection program on all remaining sentences of these debates. The topics of the sentences allow us to gain better insight into the data. The results of our study which leverages the detected topics are reported in Section 4. The high accuracy of the topic-detection program on the ground-truth data gives us confidence on the results.

3.4 Verdict Collection

We used CNN and PolitiFact as the means for comparing ClaimBuster's results. These two organizations were selected because each identifies claims they judge to be worth checking and then rates each claim on a truthfulness scale. The verdicts for CNN are true, mostly true, true but misleading, false or it's complicated. PolitiFact uses true, mostly true, half true, mostly false, false and "pants on fire" (egregiously false). Other organizations focus specifically on false or misleading claims only (Factcheck.org) or write about debate statements they found interesting or suspicious (Washington Post) which makes a comparison to ClaimBuster prob-

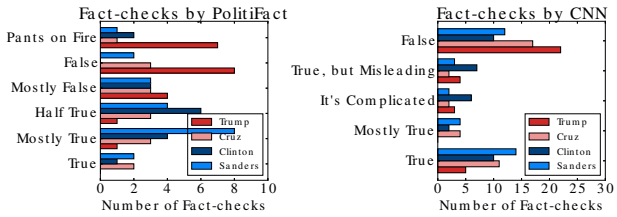

Figure 5: Distribution of verdicts for each major candidate

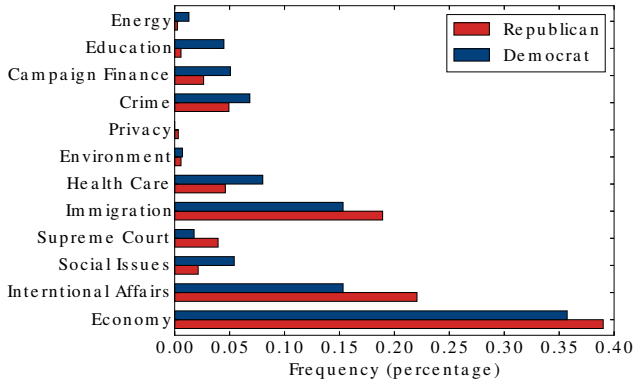

Figure 6: Distribution of topics over all the sentences for each party

lematic.

For each of the 21 debates CNN and PolitiFact prepared a summary of the factual claims they chose to check and rendered a verdict on them. We collected all of these verdicts, 224 from CNN and 118 from PolitiFact.

Table 1 shows scores given by ClaimBuster to the claims fact-checked by CNN and PolitiFact. The ClaimBuster average for sentences fact-checked by CNN is 0.433 compared to 0.258 for those sentences not selected by CNN, a statistically significant difference. Likewise, the ClaimBuster average for sentences checked by PolitiFact is 0.438 compared to 0.258 for those not selected, also a significant difference. The results of these comparisons demonstrate the utility of ClaimBuster in identifying sentences likely to contain important factual claims.

Figure 4 shows, for each party, the number of fact-checks of different veracity by CNN and PolitiFact. Figure 5 shows number of fact-checks for each major candidates. One observation is, *Donald Trump* has presented more *Pants on Fire*, *False* and *Mostly False* factual claims than other candidates according to PolitiFact. Similar observation is also evident according to CNN.

4. RESULTS

With the ClaimBuster score, topic and veracity of the sentences at hand, we study the relation among these and try to find answers to questions such as which candidate presented more factual claims pertaining to a certain topic compared to others and so on.

Figure 6 shows the distribution of topics among sentences for each party. Republican candidates are more vocal about

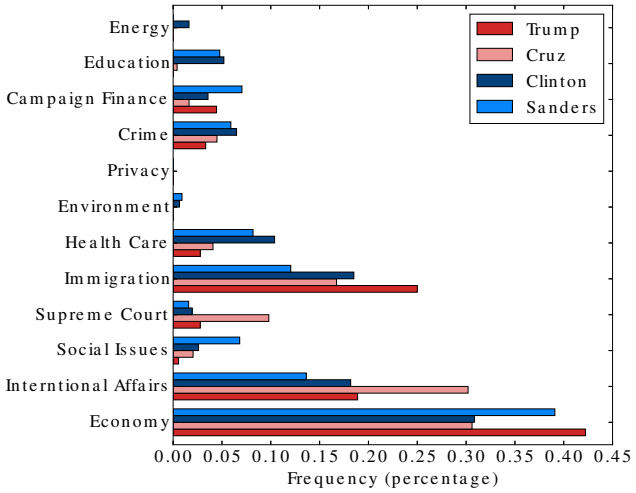

Figure 7: Distribution of topics over all the sentences from the major candidates

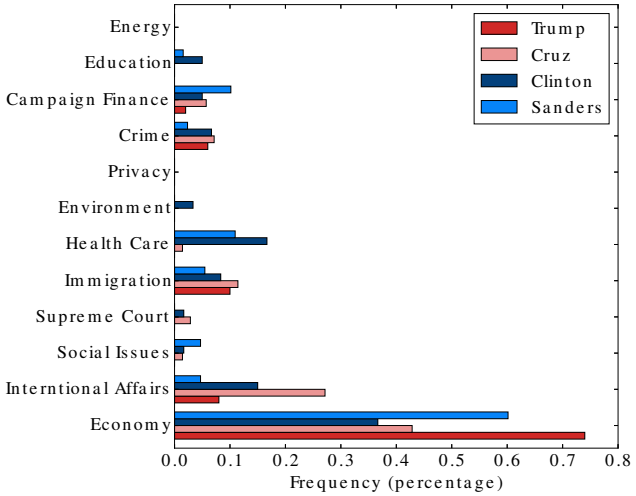

Figure 8: Distribution of topics over sentences scored high (≥ 0.5) by ClaimBuster

Economy, *International Affairs*, and *Immigration* compared to the Democrats. On the other hand, Democrats are more vocal on *Energy*, *Education*, *Social Issues* and *Health Care*. We roll down to the candidate level and try to understand the most vocal candidates on each of the topics. Figure 7 shows the topic distribution for each major candidate. *Bernie Sanders* was the most vocal on *Social Issues* among the candidates. *Ted Cruz* spoke significantly more on *International Affairs* compared to other candidates.

We analyzed the check-worthiness of the sentences of each topic. Figure 8 shows the topic distribution of sentences having ClaimBuster score ≥ 0.5. This figure explains how often the candidates used factual claims while speaking about different topics. For example, both *Donald Trump* and *Bernie Sanders* presented significantly more check-worthy factual claims relating to the *Economy* compared to their debate competitors.

Figure 9 shows the topic distribution of sentences having

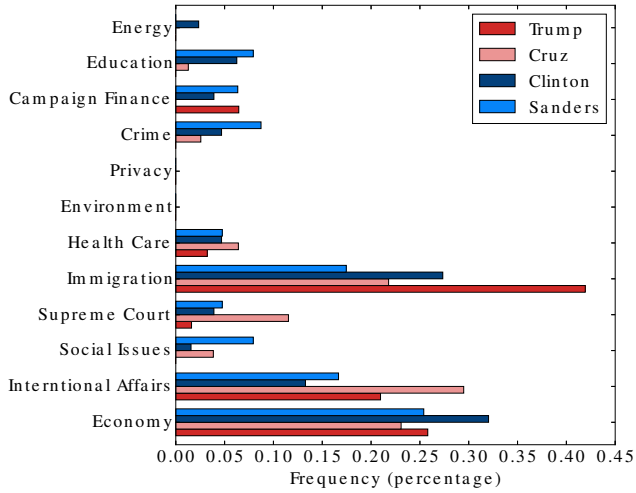

Figure 9: Distribution of topics over sentences scored low (≤ 0.3) by ClaimBuster

ClaimBuster score ≤ 0.3. This figure explains how much the candidates spoke about different topics without presenting factual claims. One interesting observation derived from Figures 8 and 9 is that Republican candidates spoke about *Health Care* but used fewer factual claims regarding this topic. On the other hand, Democratic candidate *Hillary Clinton* presented factual statements related to *Environment* rather than presenting non-factual, subjective statements.

Figure 10 shows the topic distributions of CNN, PolitiFact sentences as well as of highly check-worthy factual sentences (ClaimBuster score ≥ 0.5). This figure signifies that there are strong similarities between ClaimBuster and the fact-checking organizations. ClaimBuster tends to give high scores to the topics which CNN and PolitiFact tend to choose for fact checking. For example, all three have about 50 percent of the fact checks (or high ClaimBuster scores) associated with *Economy*, about 14 percent for *International Affairs*, about 10 percent for *Immigration* and 4 percent for *Crime*. One topic where ClaimBuster showed a difference with the human fact-checkers was *Social Issues*. That topic represented about 9 percent of the CNN and PolitiFact fact-checks but only about 2 percent of the highly scored ClaimBuster sentences.

5. WORK IN PROGRESS

We look forward to making progress on several fronts in building ClaimBuster in the future. We are applying ClaimBuster on Australian Parliament Hansard [5]. This will facilitate fact-checking statements made by the members of parliament.

Building a repository of fact-checks done by professionals is also in our agenda. This will enable automatic matching of claims during a live event with known fact-checks in the repository and instantly informing the audience about the claims' veracity.

We are also studying claims found in the media about various domains such as politics, sports and so on. Particularly,

[5]http://www.aph.gov.au/Parliamentary_Business/Hansard

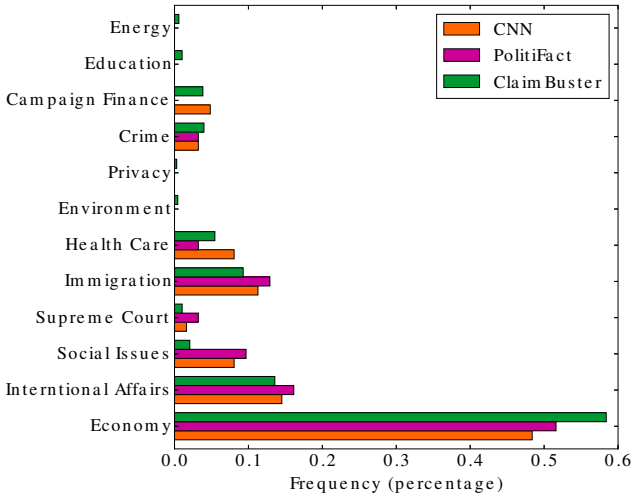

Figure 10: Comparison of topic distributions of CNN, PolitiFact fact-checked sentences and sentences scored high (≥ 0.5) by ClaimBuster

we are interested in investigating how numbers, actions and comparisons are used in factual claims. Furthermore, we will formulate templates of factual claims. Such templates will enable programs to automatically categorize claims, which can be valuable for improving fact-checking accuracy.

All these efforts will bring us closer towards the "Holy Grail" of automated fact-checking - a fully automated, live, end-to-end fact-checking system [5].

6. CONCLUSION

In this study, we used the 2016 U.S. presidential election primary debates to compare the results of our automated factual claim tool against the judgments of professional journalism organizations. Overall, we found that sentences selected by both CNN and PolitiFact for fact checking had ClaimBuster scores that were significantly higher (were more check-worthy) than sentences not selected for checking. On average, the sentences had scores nearly twice the magnitude of the unchecked sentences. At the same time, many sentences scored highly by ClaimBuster were not selected by either organization. Part of this is due to constraints; PolitiFact only checked about 6 or 7 claims per debate, CNN about 10. It may also relate to the nature of the claims, timeliness, uniqueness and other traits. These are areas for further research.

ClaimBuster also compared favorably to CNN and PolitiFact in the distribution of topics among highly scored sentences and fact-checked sentences. The percentage of sentences checked (or highly scored by ClaimBuster) were very similar for topics like *Economy*, *International Affairs*, *Immigration* and *Crime*. One difference was for social issues where ClaimBuster scores were low relative to the fact-checking judgment of CNN and PolitiFact. Understanding the discrepancies will be an area for further refinement of ClaimBuster.

7. REFERENCES

[1] P. Biyani, S. Bhatia, C. Caragea, and P. Mitra. Using non-lexical features for identifying factual and opinionative threads in online forums. *Knowledge-Based Systems*, 69:170–178, 2014.

[2] K. Fridkin, P. J. Kenney, and A. Wintersieck. Liar, liar, pants on fire: How fact-checking influences citizens' reactions to negative advertising. *Political Communication*, 32(1):127–151, 2015.

[3] J. A. Gottfried, B. W. Hardy, K. M. Winneg, and K. H. Jamieson. Did fact checking matter in the 2012 presidential campaign? *American Behavioral Scientist*, 57(11):1558–1567, 2013.

[4] L. Graves, B. Nyhan, and J. Reifler. Understanding innovations in journalistic practice: A field experiment examining motivations for fact-checking. *Journal of Communication*, 66(1):102–138, 2016.

[5] N. Hassan, B. Adair, J. T. Hamilton, C. Li, M. Tremayne, J. Yang, and C. Yu. The quest to automate fact-checking. *Proceedings of the 2015 Computation+Journalism Symposium*, 2015.

[6] N. Hassan, C. Li, and M. Tremayne. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1835–1838. ACM, 2015.

[7] M. E. McCombs and D. L. Shaw. The agenda-setting function of mass media. *Public opinion quarterly*, 36(2):176–187, 1972.

[8] B. Nyhan and J. Reifler. The effect of fact-checking on elites: A field experiment on us state legislators. *American Journal of Political Science*, 59(3):628–640, 2015.

[9] M. Schudson. The objectivity norm in american journalism. *Journalism*, 2(2):149–170, 2001.

[10] T. W. Smith. America's most important problem-a trend analysis, 1946–1976. *Public Opinion Quarterly*, 44(2):164–180, 1980.

[11] C. Spivak. The fact-checking explosion. *American Journalism Review*, 32(4):38–43, 2011.

[12] A. Vlachos and S. Riedel. Fact checking: Task definition and dataset construction. *ACL 2014*, page 18, 2014.

[13] J.-H. Zhu. Issue competition and attention distraction: A zero-sum theory of agenda-setting. *Journalism & Mass Communication Quarterly*, 69(4):825–836, 1992.